Fully Convolutional Networks for Continuous Sign Language Recognition

Ka Leong Cheng[1], Zhaoyang Yang[2], Qifeng Chen[1], and Yu-Wing Tai[1,3]

[1] The Hong Kong University of Science and Technology
{klchengad,cqf}@ust.hk
[2] Tencent
yangzhaoyang6@126.com
[3] Kwai Inc.
yuwing@gmail.com

Abstract. Continuous sign language recognition (SLR) is a challenging task that requires learning on both spatial and temporal dimensions of signing frame sequences. Most recent work accomplishes this by using CNN and RNN hybrid networks. However, training these networks is generally non-trivial, and most of them fail in learning unseen sequence patterns, causing an unsatisfactory performance for online recognition. In this paper, we propose a fully convolutional network (FCN) for online SLR to concurrently learn spatial and temporal features from weakly annotated video sequences with only sentence-level annotations given. A gloss feature enhancement (GFE) module is introduced in the proposed network to enforce better sequence alignment learning. The proposed network is end-to-end trainable without any pre-training. We conduct experiments on two large scale SLR datasets. Experiments show that our method for continuous SLR is effective and performs well in online recognition.

Keywords: Continuous sign language recognition · Fully convolutional network · Joint training · Online recognition

1 Introduction

Sign language is a common communication method for people with disabled hearing. It composes of a variety range of gestures, actions, and even facial emotions. In linguistic terms, a gloss is regarded as the unit of the sign language [27]. To sign a gloss, one may have to complete one or a series of gestures and actions. However, many glosses have very similar gestures and movements because of the richness of the vocabulary in a sign language. Also, because different people have different action speeds, a same signing gloss may have different lengths. Not to mention that different from spoken languages, sign language like ASL [22] usually does not have a standard structured grammar. These facts place additional difficulties in solving continuous SLR because it requires the model to be highly capable of learning spatial and temporal information in the signing sequences.

Early work on continuous SLR [6,18,34] utilizes hand-crafted features followed by Hidden Markov Models (HMMs) [43,48] or Dynamic Time Warping (DTW) [47] as common practices. More recent approaches achieve state-of-the-art results using CNN and RNN hybrid models [4,14,44]. However, we observe that these hybrid models tend to focus on the sequential order of seen signing sequences in the training data but not the glosses, due to the existence of RNN. So, it is sometimes hard for these trained networks to recognize unseen signing sequences with different sequential patterns. Also, training of these models is generally non-trivial, as most of them require pre-training and incorporate iterative training strategy [4], which greatly lengthens the training process. Furthermore, the robustness of previous models is limited to sentence recognition only; most of the methods fail when the test cases are signing videos of a phrase (sentence fragment) or a paragraph (several sentences). Online recognition requires good recognition responses for partial sentences, but these models usually cannot give correction recognition until the signer finishes all the signing glosses in a sentence. Such limitation in robustness makes online recognition almost impossible for CNN and RNN hybrid models.

In this paper, we propose a fully convolutional network [24] for continuous SLR to address these challenges. The proposed network can be trained end-to-end without any pre-training. On top of this, we introduce a GFE module to enhance the representativeness of features. The FCN design enables the proposed network to recognize new unseen signing sentences, or even unseen phrases and paragraphs. We conduct different sets of experiments on two public continuous SLR datasets. The major contribution of this work can be summarized:

1. We are the first to propose a fully convolutional end-to-end trainable network for continuous SLR. The proposed FCN method models the semantic structure of sign language as glosses instead of sentences. Results show that the proposed network achieves state-of-the-art accuracy on both datasets, compared with other RGB-based methods.
2. The proposed GFE module enforces additional rectified supervision and is jointly trained along with the main stream network. Compared with iterative training, joint training with the GFE module fastens the training process because joint training does not require additional fine-tuning stages.
3. The FCN architecture achieves better adaptability in more complex real-world recognition scenarios, where previous LSTM based methods would almost fail. This attribute makes the proposed network able to do online recognition and is very suitable for real-world deployment applications.

2 Related Work

There are mainly two scenarios in SLR: isolated SLR and continuous SLR. Isolated SLR mainly focuses on the scenario where glosses have been well segmented temporally. Work in the field generally solves the task with methods such as Hidden Markov Models (HMMs) [10,12,13,29,35,42], Dynamic Time Warping

(DTW) [36], and Conditional Random Field (CRF) [40,41]. As for continuous SLR, the task becomes more difficult as it aims to recognize glosses in the scenarios where no gloss segmentation is available but only sentence-level annotations as a whole. Learning separated individual glosses becomes more difficult in the weakly supervised setting. Many approaches propose to estimate the temporal boundary of different glosses first and then apply isolated SLR techniques and sequence to sequence methods [7,16] to construct the sentence.

Concerning temporal boundary estimation, Cooper and Bowden [3] develop a method to extract similar video regions for inferring alignments in videos by using data mining and head and hand tracking. Farhadi and Forsyth [8] also come up with a method that utilizes HMMs to build a discriminative model for estimating the start and end frames of the glosses in video streams with a voting method. Yin et al. [46] make further improvements by introducing a weakly supervised metric learning framework to address the inter-signer variation problem in real applications of SLR.

As for sequence to sequence methods, much work follows the framework used in the topic of speech recognition [25,33], handwriting recognition [23,32], and video captioning [39]. Specifically, an encoder module is responsible for extracting features in the input video frame sequences, and a CTC module acts as a cost function to learn the ground truth sequences. This framework also shows good performance on continuous SLR, and more recent work applies CNN and RNN hybrid models to infer gloss alignments implicitly [2,14,26,30]. However, RNNs are sometimes more sensitive to the sequential order than the spatial features. As a result, these models tend to learn much about the sequential signing patterns but little about the glosses (words), causing the failure of the recognition for unseen phrases and paragraphs.

3 Method

Formally, the proposed network aims to learn a mapping $H : \mathcal{X} \mapsto \mathcal{Y}$ that can transform an input video frame sequence $\mathcal{X}$ to a target sequence $\mathcal{Y}$. The feature extraction contains two main steps: a frame feature encoder and a two-level gloss feature encoder. On top of them, a gloss feature enhancement (GFE) module is introduced to enhance the feature learning. An overview of the proposed network is shown in Figure 1.

3.1 Main stream design

Frame feature encoder. The proposed network first encodes spatial features of the input RGB frames. The frame feature encoder S composes of a convolutional backbone S_{cnn} to extract features in the frames and a global average pooling layer S_{gap} to compress the spatial features into feature vectors. Formally, each signing sequence is a tensor with shape (t, c, h, w), where t denotes the length of the sequence, c denotes the number of channels, and h, w denotes the height

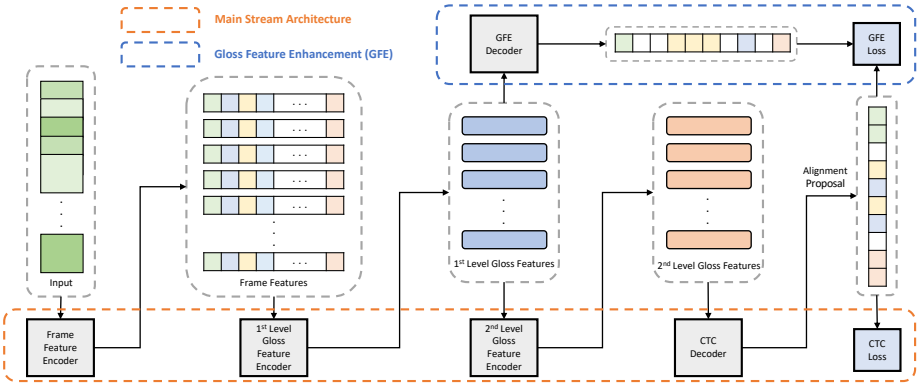

Fig. 1. Overview of the proposed network. The network is fully convolutional and divides the feature encoding process into two main steps. A GFE module is introduced to enhance the feature learning

and width of the frames. The process of encoder S can be described as:

$$\{s\}^{t\times f_s} = S(\{x\}^{t\times c\times h\times w}) = S_{gap}(S_{cnn}(\{x\}^{t\times c\times h\times w})). \tag{1}$$

The output is of shape $\{s\}^{t\times f_s}$. Note that frame feature encoder treats each frame independently for the frame (spatial) feature learning.

Two-level gloss feature encoder. The two-level gloss feature encoder G follows S immediately and aims to encode gloss features. Instead of using LSTM layers, a common practice in temporal feature encoding, we achieve this by using 1D convolutional layers over time dimension. Precisely, the first level encoder G_1 consists of 1D-CNNs with a relatively larger filter size. Pooling layers can be used between convolutional layers to increase the window size when needed. Differently, the filter size is relatively smaller for the 1D-CNNs in the second level encoder G_2, with no pooling layers used in G_2. So, G_2 does not change the temporal dimension but only reconsider the contextual information between glosses by taking into account the neighboring glosses.

The overall convolutional process of G can be interpreted as a sliding window on the frame feature vector $\{s\}^{t\times f_s}$ along the time dimension. The sliding window size l and the stride δ are determined by the accumulated receptive field size and the accumulated stride of 1D-CNNs in G_1. Let $\{g\}^{k\times f_g}$ and $\{g'\}^{k\times f_{g'}}$ be the output tensor of gloss feature encoder G_1 and G_2, respectively. The operation of the encoder G can be formulated as:

$$\{g'\}^{k\times f_{g'}} = G(\{s\}^{t\times f_s}) = G_2(G_1(\{s\}^{t\times f_s})) = G_2(\{g\}^{k\times f_g}), \tag{2}$$

where k is the number of encoded gloss features and can be calculated with:

$$k = \lfloor \frac{t-l}{\delta} \rfloor + 1. \tag{3}$$

The two-level gloss feature encoder takes into account only multiple frames at a time. The window size l should be designed to be around the average length of the signing glosses to ensure good performance during the gloss feature extraction. With a proper window size design, G_1 can better model the semantic information of a "gloss" in sign language. G_2 further considers the gloss neighborhood information to achieve better prediction.

One benefit of our FCN design over previous LSTM design is that it greatly increases the adaptability of recognition, especially for online applications. Our proposed network can provide high-quality recognition on sequences with various length, which is essential in real-world recognition scenarios. We will further discuss the advantages of the FCN design in Section 4.4.

CTC decoder. The Connectionist Temporal Classification (CTC) [9] is used as the network decoder. The CTC decoder D aims to decode the encoded gloss feature $\{g'\}^{k\times f_{g'}}$. CTC is an objective function that considers all possible underlying alignments between the input and target sequence. An extra "blank" label is added in the prediction space to match the output sequence with the target sequence in temporal dimension. Specifically, we employ a fully connected layer D_{fc} after G to cast the gloss feature dimension from $(k,\ f_{g'})$ to $(k,\ u)$ and a Softmax activation to finally transform the gloss feature to the prediction space $\{z\}^{k\times u}$:

$$\{z\}^{k\times u} = D(\{g'\}^{k\times f_{g'}}) = softmax(D_{fc}(\{g'\}^{k\times f_{g'}})), \tag{4}$$

where v is the vocabulary size and $u = v + 1$ is the size of each output with the extra "blank" added.

With normalized probabilities $\{z\}^{k\times u}$, the output alignment $\boldsymbol{\pi}$ can then be generated by taking the label with maximum likelihood at every decoding step. The final recognition result $\boldsymbol{y}$ is obtained by using the many-to-one function $\mathcal{B}$ introduced in CTC to remove repeated and blank predictions in $\boldsymbol{\pi}$. The CTC objective function is defined as the negative log-likelihood of all possible alignments matched to the ground truth:

$$\mathcal{L}_{ctc}(\boldsymbol{x}, \boldsymbol{y}) = -log\ p(\boldsymbol{y}|\boldsymbol{x}). \tag{5}$$

With the additional l_2 regularizer $\mathcal{L}_{reg}$ on the network parameters $\boldsymbol{W}$, the objective function of the main stream of the network $\mathcal{L}_{main}$ is defined as:

$$\begin{aligned}\mathcal{L}_{main} &= \mathcal{L}_{ctc} + \lambda_1 \mathcal{L}_{reg} \\ &= \frac{1}{|\mathcal{S}|} \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{S}} \mathcal{L}_{ctc}(\boldsymbol{x}, \boldsymbol{y}) + \lambda_1 ||\boldsymbol{W}||^2,\end{aligned} \tag{6}$$

where $\mathcal{S}$ is the sample space, and λ_1 is the weight factor of the regularizer.

3.2 Gloss feature enhancement

The main stream of the network has mainly two tasks: (1) alignment inference and (2) gloss prediction. The performance highly depends on how well the network can generalize on glosses, as they are the unit of sign language. Therefore,

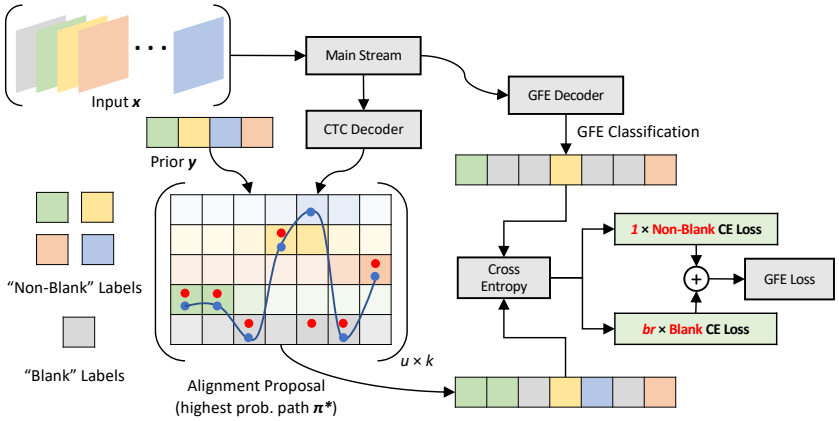

Fig. 2. The GFE module. The red dots in the prediction map are network outputs, while the blue ones are alignment proposals. The proposal rectifies the false predictions in the output to match the ground truth

it is essential to improve the quality of gloss features. Previous methods generally achieve this by incorporating iterative training strategies [4,20,31,5]. They first break training into several stages and then gradually refine feature extraction as each stage processed. However, with this strategy, whenever the training is switched to another stage, the network needs to first gradually adapt to a different objective, which greatly lengthens the number of training epochs and reduces the training efficiency. Moreover, the supervision used in some methods is generally the output of the network, which may contain some false predictions that can further reduce learning efficiency and limit the effectiveness of the refinement.

To remedy these problems, we propose the GFE module. The GFE module uses rectified supervision and is jointly trained with the main stream of the network, so it can improve the the main stream network performance on the line. We illustrate the idea of the GFE module in Figure 2.

Alignment proposal. High-quality supervision can significantly improve the effectiveness of feature enhancement. Similar to [5], we make use of the network prediction map to find a better alignment proposal as the supervision. Specifically, given an input sequence, the CTC decoder D generates a prediction map $\{z\}^{k\times u}$, which is the probability of emissions in each decoding step. Let π^* denote the element in the alignment proposal. The alignment proposal $\boldsymbol{\pi}^* = \{\pi^*\}^k$ used in the GFE module can then be generated by searching the alignment proposal with the highest probability that can be matched to the ground truth sequence:

$$\boldsymbol{\pi}^* = \underset{\boldsymbol{\pi}\in\mathcal{B}^{-1}(\boldsymbol{y})}{\arg\max}\, p(\boldsymbol{\pi}|\boldsymbol{x}), \tag{7}$$

where $\mathcal{B}^{-1}$ is the inverse function of $\mathcal{B}$. Hence, the alignment proposal is guaranteed to be a matched alignment of the ground truth sequence. Each π^* in $\boldsymbol{\pi^*}$ can be paired with a first level gloss feature vector g at the corresponding time step in $\boldsymbol{g}$, which gives a pair of learning sample $(g, \pi^*) \in \mathcal{V}$.

Joint training with weighted cross-entropy. To use the learning pairs in $\mathcal{V}$ as enhancement supervision, we add a fully connected layer F_{fc} followed by a Softmax activation after G_1. When joint training with the GFE module, gradients along this addition branch only propagate back to F and G_1 to enhance the frame and gloss feature learning. Formally, the GFE module contains a GFE decoder F, that takes $\boldsymbol{g} = \{g\}^{k \times f_g}$, the output vector of G_1, as inputs, and outputs the predicted gloss sequence in prediction space $\hat{\boldsymbol{\pi}}^* = \{\hat{\pi}^*\}^{k \times u}$:

$$\{\hat{\pi}^*\}^{k \times u} = F(\{g\}^{k \times f_g}) = softmax(F_{fc}(\{g\}^{k \times f_g})). \tag{8}$$

It is intuitive to train the gloss feature enhancement branch with cross-entropy loss. However, it is common that most of the label π^* in $\boldsymbol{\pi^*}$ is "blank" as k is generally much bigger than the number of glosses in the ground truth, causing the imbalance of samples in $\mathcal{V}$. The sample imbalance may limit the effectiveness of training for the GFE module. Therefore, we introduce a balance ratio to decrease the loss from "blank" labels. The balance ratio is defined as the proportion of "non-blank" labels in the given proposal $\boldsymbol{\pi^*}$:

$$br = \frac{\#non\text{-}blank}{\#total}. \tag{9}$$

For every $(g, \pi^*) \in \mathcal{V}$, we re-scale the cross-entropy loss to obtain the GFE loss, where the scaling factor w_i equals to br if it is blank label ($i = u$), otherwise w_i equals to 1:

$$\mathcal{L}_{gfe}(g, \pi^*) = -\frac{1}{u} \sum_{i=1}^{u} w_i log\ p(\pi = \pi_i^* | g). \tag{10}$$

With the GFE module, the overall objective of the proposed network $\mathcal{L}$ becomes:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{main} + \lambda_2 \mathcal{L}_{gfe} \\ &= \frac{1}{|\mathcal{S}|} \sum_{(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{S}} \mathcal{L}_{ctc}(\boldsymbol{x}, \boldsymbol{y}) + \\ &\quad \frac{\lambda_2}{|\mathcal{V}|} \sum_{(g, \pi^*) \in \mathcal{V}} \mathcal{L}_{gfe}(g, \pi^*) + \lambda_1 ||\boldsymbol{W}||^2, \end{aligned} \tag{11}$$

where λ_2 is the weight factor for the GFE module. Note that the network objective is unified with joint training, so the training process is more efficient.

4 Experiments

We conduct experiments on the Chinese Sign Language (CSL) dataset [14] and the RWTH-PHOENIX-Weather-2014 (RWTH) dataset [18]. We detail the experimental setup, results, ablation studies, and online recognition in this section.

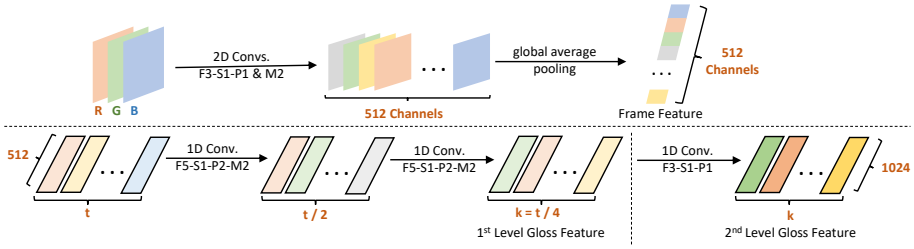

Fig. 3. The setting of the proposed main stream network. F, M refer to the filter size of convolution and max-pooling, respectively. S, P refer to the stride and padding size of convolution, respectively. Numbers aside are their actual size

4.1 Experimental setup

Dataset. The RWTH-PHOENIX-Weather-2014 (RWTH) dataset is recorded from a public weather broadcast television station in Germany. All signers wear dark clothes and perform sign languages in front of a clean background. There are 6,841 different sentences signed by 9 different signers (around 80,000 glosses with a vocabulary of size 1,232). All videos are pre-processed to a resolution of 210×260 and 25 frames per second (FPS). The dataset is officially split with 5,672 training samples, 540 validation samples, and 629 testing samples.

The Chinese Sign Language (CSL) dataset contains 100 sentences, each being signed for 5 times by 50 signers (in total 25,000 videos). Videos are shoot using a Microsoft Kinect camera with a resolution of 1280×720 and a frame rate of 30 FPS. The vocabulary size is relatively small (178); however, the dataset is richer in performance diversity, since signers wear different clothes and sign with different speeds and action ranges. With no official split given, we divide the dataset into a training set of 20,000 samples and a testing set of 5,000 samples and ensure that the sentences in the training and testing sets are the same, but the signers are different.

Evaluation metric. We use word error rate (WER), which is the metric commonly used in continuous SLR, to evaluate the performance of recognition:

$$WER(H(\boldsymbol{x}), \boldsymbol{y})) = \frac{\#ins + \#del + \#sub}{\#labels\ in\ \boldsymbol{y}}. \tag{12}$$

We treat a Chinese character as a word during evaluation for the CSL dataset.

Implementation details. The main stream network setting used in our experiments is shown in Figure 3. For S_{cnn}, the channel number gradually increases in this pattern: 3-32-64-64-128-128-256-256-512-512. F3-S1-P1 is used in each layer, and an additional M2 is added if the channel number increases. S_{gap} does global average pooling on each channel, so each frame is encoded as an array

with a length of 512. For encoder G_1, two F5-S1-P2-M2 layers are used, and the channel number remains unchanged as 512. For encoder G_2, one F3-S1-P1 layer is used, and the channel number increases to 1024. Both fully connected layers D_{fc} and F_{fc} in the main stream and the GFE module cast the input channel number to u, the number of vocabulary size plus one blank label.

Batch normalization [15] is added after every convolutional layer to accelerate training. The input resolution of the network is 224×224. The window size in the first level gloss feature encoder is set to be 16 (about 0.5-0.6 seconds), which is the average time needed for completing a gloss, and the stride of the window is set to be 4. The second level gloss feature encoder further considers 3 adjacent gloss features for better prediction.

We use Adam [17] optimizer for training. We set the initial learning rate to be 10^{-4}. The weight factor λ_1 and λ_2 are empirically set to be 10^{-4} and 0.05, respectively. For the RWTH dataset, we train the proposed network for 80 epochs and halve learning rate at epoch 40 and 60. For the CSL dataset, the network is trained for 60 epochs, with the learning rate reduced by half at epoch 30 and 45. For data augmentation, all frames are first resized to 256×256 and then randomly cropped to fit the input shape. We also do temporal augmentation by first scaling up the sequence by $+20\%$ and then by -20%. Joint training with the GFE module is activated after epoch 15 for RWTH and epoch 10 for CSL, which are chosen through experiments to avoid unreliable alignment proposal at the initial optimization stage. The alignment proposals in the GFE module are updated every 10 epochs. When updating the proposal, temporal augmentation is disabled.

4.2 Results

We give a thorough comparison between the proposed network and previous RGB-based methods on both datasets. The results of previous methods are collected from their original papers. Please note that we mainly focus on online recognition in SLR, where the inputs are usually RGB video frames. Hence, we only compare our results with previous methods that use solely RGB modality.

Results on the CSL dataset and the RWTH dataset are shown in Table 1 and Table 2, respectively. We see that the proposed network achieves state-of-the-art performance on both datasets for RGB-based methods. The best result achieves 3.0% for the CSL dataset. For the RWTH dataset, our model reports 23.7% for the development set and 23.9% for the testing set. We also test the performance on the recognition partition (without translation) of the RWTH-PHOENIX Weather 2014 **T** dataset [1]. Our WER on the development set and testing set are 23.3% and 25.1%, respectively. These results illustrate the effectiveness of the proposed network.

To further demonstrate the effectiveness of our methods, we show some sample outputs in Figure 4 to compare the recognition quality of different network settings. We observe that in the LSTM setting, models recognize the identical glosses (such as "OST" in sample 1 and "HIER" in sample 3) in a sentence as different words, and that errors usually occur adjacently in the sentences.

Table 1. Result comparison on CSL

Methods	WER
DTW-HMM [47]	28.4
LSTM [38]	26.4
S2VT [37]	25.5
LSTM-A [45]	24.3
LSTM-E [28]	23.2
HAN [43]	20.7
LS-HAN [14]	17.3
SubUNet [2]	11.0
HLSTM [11]	7.6
HLSTM-attn [11]	7.1
Align-iOpt [31]	6.1
SF-Net [44]	3.8
Ours	**3.0**

Table 2. Result comparison on RWTH

Methods	WER	
	Dev	Test
Koller et al. [18]	57.3	55.6
Deep Hand [19]	47.1	45.1
Deep Sign [21]	38.3	38.8
SubUNet [2]	40.8	40.7
Cui et al. [4]	39.4	38.7
LS-HAN [14]	-	38.3
Align-iOpt [31]	37.1	36.7
SF-Net [44]	38.0	38.1
Re-Sign [20]	27.1	26.8
STMC (RGB) [49]	25.0	-
Cui et al. (RGB) [5]	23.8	24.4
Ours	**23.7**	**23.9**

In contrast, the proposed network produces consistent results for the identical glosses in a sentence, and errors are usually isolated. The observations infer that the LSTM based methods tend to learn robust sequential information, while the proposed network focuses on learning strong gloss features. So, we claim that the proposed network has a better generalization capability, because identical glosses are consistently classified, and errors do not have significant effects on neighboring glosses. On top of that, the GFE module further improves the performance by rectifying wrong recognition (such as "IN-KOMMEND" in sample 3), finding missing recognition (such as "AUCH" in sample 2), and adjusting alignments. More qualitative comparison can be found in the supplementary material.

4.3 Ablation studies

In this section, we present further ablation studies to demonstrate the effectiveness of our method.

Temporal feature encoder. We first conduct a set of experiments to compare network performance with different temporal feature encoder designs. We test 6 different design combinations for the temporal feature encoder. For notation, **None** means no architecture; **LSTM** or **BiLSTM** means 1 LSTM or BiLSTM layer with 512 hidden states, respectively; **1D-CNN** means two F5-S1-P2-M2 1DConvs for the 1^{st} level or one F3-S1-P1 1DConv for the 2^{nd} level. We show results on the testing set of the RWTH and CSL datasets in Table 3. Note that in this set of experiments, the GFE module is not activated.

We see that both levels of gloss feature encoder are essential for recognition as WER raises significantly when either of them is absent. CNN-based designs consistently outperform their LSTM counterparts for the RWTH dataset, but networks with BiLSTM give the best results for the CSL dataset. We should

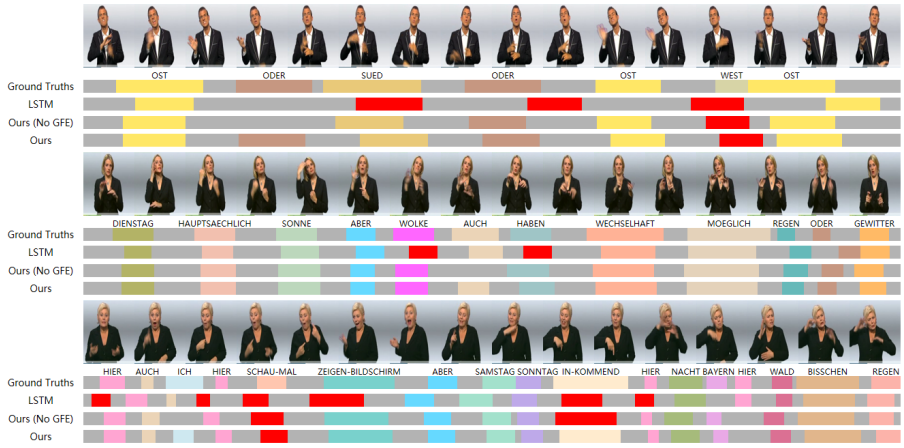

Fig. 4. Sample outputs for different network settings. Wrong recognitions (except deletion errors) are highlighted in red. Ground truths are manually aligned

be reminded of the differences between the two datasets. The CSL dataset has richer diversity in the spatial dimension (such as different cloth colors) than in the temporal term. In other words, it is easier for the network to learn the temporal features than the spatial features in the CSL dataset. Also, different from that all the testing sentences are unseen in the RWTH training set, all the testing sentences in the CSL dataset are already seen in the training set, but just signed by different people. The working nature of 1D-CNN and LSTM is also different. The LSTM based method has direct access to sentence-level information, since LSTM layers have access to all the frame information. While the FCN method has less sentence-level supervision (only indirect access through the CTC decoding function), as the FCN model uses only a fixed number of frames to predict a gloss at each time step with 1D-CNNs.

Table 3. Network performance with different temporal feature encoder design

1^{st} level	2^{nd} level	WER	
		RWTH	CSL
None	**1D-CNN**	60.5	23.3
1D-CNN	**None**	42.1	10.4
LSTM	**1D-CNN**	32.1	10.8
LSTM	**BiLSTM**	30.8	3.6
1D-CNN	**BiLSTM**	26.5	**3.4**
1D-CNN	**1D-CNN**	**26.0**	8.2

Table 4. Network performance in different GFE module settings, br refers to the balance ratio

	GFE	br	WER
RWTH	✗	✗	26.0
	✓	✗	25.4
	✓	✓	**23.9**
CSL	✗	✗	8.2
	✓	✗	4.5
	✓	✓	**3.0**

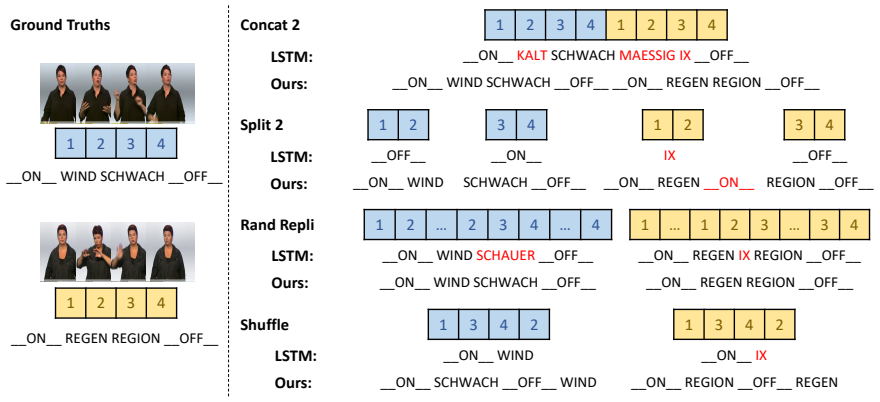

Fig. 5. Two samples are chosen for illustrating four different simulating scenarios for real-world recognition. Given that both the LSTM based method and the proposed network can recognize the original samples correctly, in all the simulating scenarios, the LSTM based method provides many false predictions while ours can preserve its accuracy. Errors are marked in red

Therefore, we claim that LSTM based methods tend to "remember" all the signing sequences in the training set instead of trying to learn the glosses independently. When testing the sentences in the CSL dataset, whose sequential sentences are the same as the training samples, the strong sequential information learned by LSTM based methods is significant and helpful, causing a relatively low reported WER. While for our FCN method, it is hard to fully extract the spatial and temporal features with weak supervision without substantial sentence-level information. Thus, it is essential to have a GFE module for feature enhancement, and the full version of our proposed network outperforms all the LSTM based methods for both datasets.

GFE module. We conduct a set of experiments to investigate the effectiveness of the GFE module. We test different settings on both the RWTH and CSL datasets, including learning without the GFE module, with the GFE module but without balance ratio, and with the GFE module and balance ratio. Results on the testing sets are shown in Table 4.

We see that when the balance ratio is used, the GFE module can significantly fine-tune the features and improve the performances accordingly for both datasets. The GFE module with balance ratio improves the testing WER for the RWTH dataset by 2.1%; the improvement is more prominent for the CSL dataset by 5.2%. The difference of improvement is because the CSL dataset has much richer diversity in the spatial dimension than the RWTH dataset, making the spatial features in the CSL dataset more difficult to be learned without the GFE module.

4.4 Online recognition

We mention in Section 4.3 that the proposed network has weaker supervision on sentence information. The FCN method focuses more on glosses rather than sentences, which directs us for some interesting experiments with different setups.

Simulating experiments. For recognition in the real world, it is natural to consider the following three scenarios: (1) Signers sign several sentences at a time. (2) Signers sign only a phrase at a time. (3) Signers may pause for some while (stutters of actions) in the middle of signing. Accordingly, as shown in Figure 5, we design three types of experiments which are conducted on the RWTH dataset to simulate online recognition: (1) concatenate multiple samples for a new sample (numbers after **Concat** indicate the number of samples being concatenated); (2) evenly split a sample into multiple samples (numbers after **Split** indicate the number of equal segments); (3) randomly select 5 frames in each sample and replace them with 12 replications in place (**Rand Repli** means random replication).

The LSTM based method may have too strong supervision of the sequential order, making it sensitive to gloss order. But one advantage of the proposed network is that the FCN model learns the glosses independently, so it is more robust to order-independent representations. To show this advantage, we may experimentally by shuffling the glosses in the testing samples. However, given no isolated annotations for individual glosses, we cannot manually construct "new" sentences with different signing orders by random shuffling. To mimic the shuffling idea, our fourth experiment is an imperfect but reasonable gloss shuffling experiment, as shown as **Shuffle** in Figure 5. We first temporally segment the input into two equal parts and insert one into another.

The results of the four experiments are shown in Table 5. It is observed that the performance of the LSTM based network degrades dramatically in all these four types of simulating scenarios. On the contrary, the proposed network shows consistent overall performance across different scenarios. We only observe additional minor errors in the output steps where samples are combined or split due to the action inconsistencies introduced in boundary places.

Table 5. Network performances in different real-world simulating scenarios on RWTH

Setups	Dev		Test	
	LSTM	Ours	LSTM	Ours
Original	31.7	23.7	30.8	23.9
Split 2	45.6	26.6	42.3	26.4
Split 3	50.2	28.0	45.3	27.6
Concat 2	40.5	24.3	41.1	24.9
Concat 3	46.0	24.8	45.7	25.0
Concat All	-	25.5	-	25.3
Rand Repli	39.1	24.6	39.5	24.6
Shuffle	58.9	27.3	55.2	28.1

Discussion. Considering the nature of the FCN design, the results further inspire us that even the proposed network is continuously being fed only a few frames that are needed (adequate) to infer the output, it can still combine all the intermediate outputs to give the same final recognition result. We use this technique to test for the **Concat All** scenario, where all testing samples are concatenated together as one large sample. Unfortunately, the LSTM based model fails to provide any result for the **Concat All** scenario, as the memory capacity limits the network to take such a large sample as an input.

All the results Table 5 indicate that the proposed network has more generalization capability and better flexibility for recognition in complex real-world recognition scenarios. The FCN design enables the proposed network to significantly reduce the memory usage for recognition. Meanwhile, indicated by the results in the **Split** and **Concat** scenarios, besides recognizing signing sentences, our method gives accurate recognition results for signing phrases and paragraphs. We can further conclude from the results in the **Split** scenarios that there is no need to wait for the arrival of all the signing glosses during the recognition process, as accurate intermediate (partial) recognition results can be given whenever adequate frames are available to the proposed network. With this great property, our method can provide intermediate results word by word along time, which is very friendly from a human-computer interaction perspective for SLR users. These properties make the proposed network have a promising application prospect for online recognition. A visual demonstration is shown in the supplementary demo video.

5 Conclusions

In this paper, we are the first to propose a fully convolutional network that can be trained end-to-end without pre-training for continuous SLR. A jointly trained GFE module is introduced to enhance the representativeness of features. Experimental results show that the proposed network achieves state-of-the-art performance on benchmark datasets with RGB-based methods. For recognition in real-world scenarios where the LSTM based network mostly fails, the proposed network reaches consistent performance and shows many great advantageous properties. These advantages make our proposed network robust enough to do online recognition.

One possible future research direction for continuous SLR is to strengthen the supervision by using the fact that some glosses are combinations of letter signs; however, this may require additional labeling pre-processing and professional knowledge in sign language. Also, the better gloss recognition accuracy obtained by the proposed network may have a good research prospect in sign language translation (SLT). Furthermore, we hope the proposed network can inspire future studies on sequence recognition tasks to investigate FCN architecture as an alternative to LSTM based models, especially for those tasks with limited data for training.

References

1. Camgoz, N., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 7784–7793 (2018) 9
2. Camgoz, N.C., Hadfield, S., Koller, O., Bowden, R.: Subunets: End-to-end hand shape and continuous sign language recognition. In: Proceedings of IEEE International Conference on Computer Vision. pp. 3075–3084 (2017) 3, 10
3. Cooper, H., Bowden, R.: Learning signs from subtitles: A weakly supervised approach to sign language recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 2568–2574 (2009) 3
4. Cui, R., Liu, H., Zhang, C.: Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 1610–1618 (2017) 2, 6, 10
5. Cui, R., Liu, H., Zhang, C.: A deep neural framework for continuous sign language recognition by iterative training. IEEE Transactions on Multimedia **21**, 1880–1891 (2019) 6, 10
6. Evangelidis, G.D., Singh, G., Horaud, R.: Continuous gesture recognition from articulated poses. In: Proceedings of European Conference on Computer Vision. pp. 595–607 (2015) 2
7. Fang, G., Gao, W.: A srn/hmm system for signer-independent continuous sign language recognition. In: Proceedings of IEEE International Conference on Automatic Face Gesture Recognition. pp. 312–317 (2002) 3
8. Farhadi, A., Forsyth, D.: Aligning asl for statistical translation using a discriminative word model. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 1471–1476 (2006) 3
9. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of International Conference on Machine Learning. pp. 369–376 (2006) 5
10. Guo, D., Zhou, W., Li, H., Wang, M.: Online early-late fusion based on adaptive hmm for sign language recognition. ACM Transactions on Multimedia Computing, Communications, and Applications **14**, 1–18 (2017) 2
11. Guo, D., Zhou, W., Li, H., Wang, M.: Hierarchical lstm for sign language translation. In: Proceedings of AAAI Conference on Artificial Intelligence. pp. 6845–6852 (2018) 10
12. Guo, D., Zhou, W., Wang, M., Li, H.: Sign language recognition based on adaptive hmms with data augmentation. In: Proceedings of IEEE International Conference on Image Processing. pp. 2876–2880 (2016) 2
13. Han, J., Awad, G., Sutherland, A.: Modelling and segmenting subunits for sign language recognition based on hand motion analysis. Pattern Recognition Letters **30**, 623–633 (2009) 2
14. Huang, J., Zhou, W., Zhang, Q., Li, H., Li, W.: Video-based sign language recognition without temporal segmentation. In: Proceedings of AAAI Conference on Artificial Intelligence. pp. 2257–2264 (2018) 2, 3, 7, 10
15. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of International Conference on Machine Learning. pp. 448–456 (2015) 9

16. Kelly, D., McDonald, J., Markham, C.: Recognizing spatiotemporal gestures and movement epenthesis in sign language. In: Proceedings of IEEE International Conference on Image Processing and Machine Vision. pp. 145–150 (2009) 3
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR preprint CoRR:1412.6980 (2014) 9
18. Koller, O., Forster, J., Ney, H.: Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. Computer Vision and Image Understanding **141**, 108–125 (2015) 2, 7, 10
19. Koller, O., Ney, H., Bowden, R.: Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 3793–3802 (2016) 10
20. Koller, O., Zargaran, S., Ney, H.: Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3416–3424 (2017) 6, 10
21. Koller, O., Zargaran, S., Ney, H., Bowden, R.: Deep sign: Hybrid cnn-hmm for continuous sign language recognition. In: Proceedings of British Machine Vision Conference. pp. 136.1–136.12 (2016) 10
22. Liddell, S.K.: Grammar, gestures, and meaning in american sign language. Cambridge: Cambridge University Press. pp. 52–53 (2003) 1
23. Liwicki, M., Graves, A., Bunke, H., Schmidhuber, J.: A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks. In: Proceedings of International Conference on Document Analysis and Recognition. pp. 367–371 (2007) 3
24. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440 (2015) 2
25. Miao, Y., Gowayyed, M., Metze, F.: Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding. In: IEEE Conference on Automatic Speech Recognition and Understanding Workshops. pp. 167–174 (2015) 3
26. Molchanov, P., Yang, X., Gupta, S., Kim, K., Tyree, S., Kautz, J.: Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 4207–4215 (2016) 3
27. Ong, S., Ranganath, S.: Automatic sign language analysis: A survey and the future beyond lexical meaning. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**, 873–91 (2005) 1
28. Pan, Y., Mei, T., Yao, T., Li, H., Rui, Y.: Jointly modeling embedding and translation to bridge video and language. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 4594–4602 (2015) 10
29. Pitsikalis, V., Theodorakis, S., Vogler, C., Maragos, P.: Advances in phonetics-based sub-unit modeling for transcription alignment and sign language recognition. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1–6 (2011) 2
30. Pu, J., Zhou, W., Li, H.: Dilated convolutional network with iterative optimization for continuous sign language recognition. In: Proceedings of International Joint Conference on Artificial Intelligence. pp. 885–891 (2018) 3
31. Pu, J., Zhou, W., Li, H.: Iterative alignment network for continuous sign language recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 4165–4174 (2019) 6, 10

32. Puigcerver, J.: Are multidimensional recurrent layers really necessary for handwritten text recognition? In: Proceedings of International Conference on Document Analysis and Recognition. pp. 67–72 (2017) 3
33. Sak, H., Senior, A., Rao, K., İrsoy, O., Graves, A., Beaufays, F., Schalkwyk, J.: Learning acoustic frame labeling for speech recognition with recurrent neural networks. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 4280–4284 (2015) 3
34. Sun, C., Zhang, T., Bao, B.K., Xu, C., Mei, T.: Discriminative exemplar coding for sign language recognition with kinect. IEEE Transactions on Cybernetics **43**, 1418–1428 (2013) 2
35. Theodorakis, S., Katsamanis, A., Maragos, P.: Product-hmms for automatic sign language recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 1601–1604 (2009) 2
36. Vela, A.H., Bautista, M., Perez-Sala, X., Ponce-López, V., Escalera, S., Baró, X., Pujol, O., Angulo, C.: Probability-based dynamic time warping and bag-of-visual-and-depth-words for human gesture recognition in rgb-d. Pattern Recognition Letters **50**, 112–121 (2014) 3
37. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence – video to text. In: Proceedings of IEEE International Conference on Computer Vision. pp. 4534–4542 (2015) 10
38. Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., Saenko, K.: Translating videos to natural language using deep recurrent neural networks. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1494–1504 (2015) 10
39. Wang, B., Ma, L., Zhang, W., Liu, W.: Reconstruction network for video captioning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 7622–7631 (2018) 3
40. Yang, H.D., Lee, S.W.: Robust sign language recognition with hierarchical conditional random fields. In: Proceedings of IEEE International Conference on Pattern Recognition. pp. 2202–2205 (2010) 3
41. Yang, R., Sarkar, S.: Detecting coarticulation in sign language using conditional random fields. In: Proceedings of IEEE International Conference on Pattern Recognition. pp. 108–112 (2006) 3
42. Yang, R., Sarkar, S.: Gesture recognition using hidden markov models from fragmented observations. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 766–773 (2006) 2
43. Yang, W., Tao, J., Ye, Z.: Continuous sign language recognition using level building based on fast hidden markov model. Pattern Recognition Letters **78**, 28–35 (2016) 2, 10
44. Yang, Z., Shi, Z., Shen, X., Tai, Y.W.: Sf-net: Structured feature network for continuous sign language recognition. arXiv preprint arXiv:1908.01341 (2019) 2, 10
45. Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C.J., Larochelle, H., Courville, A.C.: Describing videos by exploiting temporal structure. In: Proceedings of IEEE International Conference on Computer Vision. pp. 4507–4515 (2015) 10
46. Yin, F., Chai, X., Zhou, Y., Chen, X.: Weakly supervised metric learning towards signer adaptation for sign language recognition. In: Proceedings of British Machine Vision Conference. pp. 35.1–35.12 (2015) 3

47. Zhang, J., Zhou, W., Li, H.: A threshold-based hmm-dtw approach for continuous sign language recognition. In: Proceedings of International Conference on Internet Multimedia Computing and Service. pp. 237–240 (2014) 2, 10
48. Zhang, J., Zhou, W., Xie, C., Pu, J., Li, H.: Chinese sign language recognition with adaptive hmm. In: Proceedings of IEEE International Conference on Multimedia and Expo. pp. 1–6 (2016) 2
49. Zhou, H., Zhou, W., Zhou, Y., Li, H.: Spatial-temporal multi-cue network for continuous sign language recognition. In: Proceedings of AAAI Conference on Artificial Intelligence. pp. 13009–13016 (2020) 10